Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 9, No. 2, Juni 2024 (79-89) **https://doi.org/10.32493/informatika.v9i2.38971**

Implementasi Teks Mining Pada Website Kemenkes Dengan Metode LDA Menggunakan Algoritma K-Means

Ari Setiawan[1]*, Deden Wahiddin[2], Cici Emilia Sukmawati[3]

[123]Teknik Informatika, Universitas Buana Perjuangan Karawang, Jl. HS. Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat 41361
e-mail: [1]if20.arisetiawan@mhs.ubpkarawang.ac.id, [2]deden.wahiddin@ubpkarawang.ac.id, [3]cici.emilia@ubpkarawang.ac.id
*Corresponding author

Submitted Date: April 26th, 2024 Reviewed Date: Mei 10th, 2024
Revised Date: June 08th, 2024 Accepted Date: June 30th, 2024

Abstract

This research aims to improving the accessibility and management of health information on the Ministry of Health (Kemenkes) website. Before this research was conducted, content on the Ministry of Health's website was scattered without a clear structure, making it difficult for users to find the health information they needed quickly and efficiently. This results in a decrease in the quality of the user experience and a potential decrease in trust in official health information sources. With the aim of making it easier for users to find relevant information, this research uses the K-Means algorithm to group website content based on themes. Through the text mining method, five main clusters were identified, covering topics such as emergency health, certain diseases, and innovations in handling COVID-19. The results show an increase in navigation efficiency with clustering accuracy reaching 72%. The conclusion of this research is that this grouping succeeded in improving the structure and quality of information on the Ministry of Health's website, supporting data-based decision making, and improving public health services.

Keywords: text mining, topic modeling, LDA model, K-Means

Abstrak

Penelitian ini bertujuan meningkatkan aksesibilitas dan pengelolaan informasi kesehatan di situs web Kementerian Kesehatan (Kemenkes). Sebelum penelitian ini dilakukan, konten pada situs web Kemenkes tersebar tanpa struktur yang jelas, membuat pengguna kesulitan menemukan informasi kesehatan yang mereka butuhkan dengan cepat dan efisien. Hal ini mengakibatkan penurunan kualitas pengalaman pengguna dan potensi penurunan kepercayaan terhadap sumber informasi kesehatan resmi. Dengan tujuan mempermudah pengguna dalam menemukan informasi relevan, penelitian ini menggunakan algoritma K-Means untuk mengelompokkan konten situs web berdasarkan tema. Melalui metode text mining, lima klaster utama berhasil diidentifikasi, mencakup topik seperti kesehatan darurat, penyakit tertentu, dan inovasi penanganan COVID-19. Hasil menunjukkan peningkatan efisiensi navigasi dengan akurasi klasterisasi mencapai 72%. Simpulan dari penelitian ini adalah bahwa pengelompokan ini berhasil meningkatkan struktur dan kualitas informasi di situs web Kemenkes, mendukung pengambilan keputusan berbasis data, dan meningkatkan layanan kesehatan masyarakat.

Kata kunci: teks mining, topik modeling, LDA model, K-Means

1. Pendahuluan

Kemajuan *teknologi* informasi dan internet memungkinkan akses *global* pada informasi, termasuk munculnya informasi secara *online* (Doni 2022). Informasi digital terbaru disebarkan melalui berbagai saluran komunikasi dengan fokus pada keakuratan, ketertarikan, dan relevansinya bagi mayoritas orang (Woro 2022). Pesan ini menggambarkan bahwa informasi terkini disebarkan melalui platform digital, seperti

website, media sosial dan *online news* (Wahyudin 2020).

Informasi digital telah menjadi sumber informasi utama bagi pengguna internet dalam mencari informasi yang penting, sehingga memungkinkan berbagai media dan instansi memiliki website informasi berita sendiri (Wahyudin 2020). Informasi secara digital fokus pada kecepatan dan ketepatan dalam menyajikan berita dari berbagai tingkat, serta mengidentifikasi dan menyorot isu-isu terpopuler untuk menarik minat pembaca (Rusdhi and Sari 2022). Website kemenkes berperan sebagai saluran digital untuk menyajikan informasi kesehatan lokal dan global secara transparan, meski teknologi membawa tantangan dalam penyajian dan pengelolaan topik yang relevan (Anjar, Ritonga, and Toni 2021). Begitu banyaknya website yang menyajikan informasi-informasi secara *online* yang memberikan informasi secara berlebihan dalam mempresentasikan topik sehingga membuat kesalahan dalam penyampaian informasi.

Analisis teks menggunakan topik modeling adalah cara mengelompokkan teks ke dalam topik khusus dengan pendekatan klastering dalam machine learning untuk memudahkan pengelompokkan berdasarkan kesamaan topik (Nurlayli and Nasichuddin 2019). Blei dan Jordan memperkenalkan *Latent Dirichlet Allocation (LDA)* sebagai salah satu teknik utama dalam pemodelan topik (Cahyono and Angga Reni Dwi Astuti 2023). LDA adalah metode pembelajaran mesin tanpa pengawasan yang mengelompokkan data teks besar untuk mengungkap variabel tersembunyi dengan model probabilitas generatif dan *analisis hierarki bayesian* (Dinda Adimanggala, Fitra Abdurrachman Bachtiar, and Eko Setiawan 2021).

Studi sebelumnya Nurlayli dan Nasichuddin (2019) menerapkan model pemodelan topik pada data dari *Google Scholar*, menggunakan nilai kohesi 0,4076 untuk mengevaluasi 1-9 topik. Hasilnya menunjukkan bahwa judul penelitian dosen JPTEI UNY dikelompokkan ke dalam 4 klaster: pendidikan vokasi, pengembangan sistem, media pembelajaran, dan sistem pembelajaran di SMK (Nurlayli and Nasichuddin 2019).

Dalam studi terbaru oleh Bangkalang (2023) berjudul "Penerapan *Topic Modeling* pada Judul Tugas Akhir Mahasiswa dengan Metode LDA," data dari perpustakaan Universitas Kristen Satya Wacana tahun 2015-2021 digunakan untuk menghasilkan klaster topik seperti *technology acceptance model*, *framework cobit*, dan *arsitektur enterprise* dengan *coherence score* 0,617789, memberikan referensi bagi judul tugas akhir (Bangkalang 2023).

Penelitian Patmawati, dan Yusuf (2021) mengulas penggunaan Twitter oleh pejabat negara, terutama akun Presiden Jokowi, dengan menerapkan metode *Latent Dirichlet Allocation (LDA)* pada data yang diambil dari Twitter. Evaluasi model menunjukkan *perplexity* -8.069 dan *coherence score* 0.375. Hasil penelitian menyoroti bahwa tweet Presiden Jokowi menggaris bawahi fokusnya pada COVID-19 dan vaksinasi di Indonesia (Patmawati and Yusuf 2021).

Penelitian oleh Santoso et al. (2022) "Analisis Topik Tagar Covid Indonesia pada Instagram Menggunakan *Latent Dirichlet Allocation (LDA)*", dataset Instagram digunakan dengan metode LDA. Hasil pengujian menunjukkan *perplexity* -8.501 dan topic *coherence* 0.554, mengidentifikasi topik-topik seputar COVID-19, keramaian, kesehatan, dan PPKM (Santoso et al. 2022).

Pada penelitian sebelumnya belum ditemukan adanya implementasi evaluasi menggunakan algoritma klastering pada hasil pemodelan topik, dalam hal ini masih melakukan evaluasi seperti *kohesi, coherence score*, dan *topic coherence*.

Berdasarkan dari hasil penelitian sebelumnya maka akan digunakan algoritma *Latent Dirichlet Allocation (LDA)* untuk melakukan topik modeling pada laman berita kementerian kesehatan Indonesia, dan akan membandingkan performa algoritma K-Means dan k-medoids dengan menggunakan *silhouette scores*. Dalam melakukan hasil evaluasi menggunakan *elbow metode*, *Calinski Harabasz Score dan Sum of Squares Error (SSE).*

Kami memilih algoritma K-Means untuk pemodelan teks dalam analisis konten website Kementerian Kesehatan (Kemenkes) karena algoritma ini efektif untuk mengelompokkan dokumen teks berdasarkan kesamaan fitur. K-Means bekerja dengan meminimalkan jarak antara teks dan pusat cluster, sehingga menghasilkan kelompok yang homogen di dalam dan heterogen antar kelompok.

K-Means sangat cocok untuk analisis teks karena dapat mengelompokkan artikel atau dokumen berdasarkan tema atau topik yang serupa, seperti informasi penyakit, program kesehatan, dan kebijakan publik. Algoritma ini mampu menangani

data teks dalam dimensi tinggi dengan menggunakan representasi vektor teks yang tepat.

Penelitian oleh Smith et al. (2023) menunjukkan bahwa K-Means efektif dalam mengelompokkan artikel berita kesehatan berdasarkan topik, yang membantu dalam penanganan informasi kesehatan yang lebih terstruktur.

Menurut studi oleh Johnson dan Wang (2022), K-Means dapat digunakan bersama teknik representasi teks modern seperti BERT untuk meningkatkan akurasi dalam analisis konten Kesehatan.

Penyebaran informasi kesehatan yang efektif dan efisien adalah salah satu tugas utama Kementerian Kesehatan (Kemenkes). Dengan banyaknya konten yang dipublikasikan di situs web Kemenkes, seperti informasi mengenai penyakit, program kesehatan, kebijakan, dan edukasi kesehatan, pengelolaan dan penyusunan informasi ini menjadi tantangan yang signifikan. Untuk menghadapi tantangan ini, diperlukan metode yang mampu mengelompokkan dan menganalisis konten secara otomatis dan akurat.

Algoritma K-Means dapat digunakan untuk mengelompokkan konten situs web Kemenkes berdasarkan tema yang serupa. Ini mempermudah navigasi pengguna dan membantu pejabat menganalisis kebijakan serta program kesehatan secara lebih efektif.

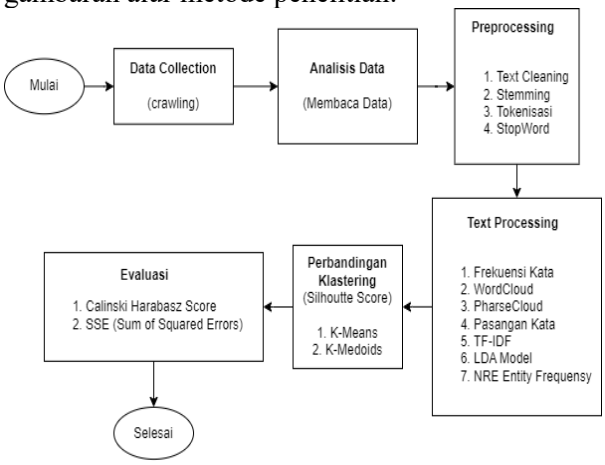
Penelitian ini bertujuan untuk menerapkan algoritma K-Means dalam pemodelan teks untuk mengelompokkan konten di situs web Kemenkes. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengelolaan informasi kesehatan, mendukung pengambilan keputusan yang lebih baik, dan meningkatkan aksesibilitas informasi kesehatan bagi publik.

Dengan adanya analisis dan pengelompokan konten yang lebih baik, Kemenkes dapat memastikan bahwa informasi kesehatan yang penting dan relevan lebih mudah ditemukan, yang pada akhirnya dapat mendukung upaya peningkatan kesehatan masyarakat secara keseluruhan.

2. Metode Penelitian

Penelitian ini berfokus pada berita yang berkaitan dengan kesehatan dan memanfaatkan informasi dari laman website kemenkes Indonesia. Data tersebut peneliti peroleh dari link berikut: https://sehatnegeriku.kemkes.go.id.

Penelitian dimulai dengan menelusuri penelitian terdahulu tentang topik modeling, *Latent Dirichlet Allocation (LDA)*, dan algoritma klastering sebagai sumber informasi yang dapat dipercaya. Kemudian, data diambil dari laman berita kementerian kesehatan Indonesia. Berikut gambaran alur metode penelitian:



Gambar 1. Alur Metode Penelitian

2.1 *Text Mining*

Text mining adalah analisis data teks yang menggunakan perangkat lunak khusus untuk menemukan konsep, pola, topik, kata kunci, dan atribut lain dalam dataset besar yang umumnya tak terstruktur (Indrayuni 2019). *Text mining* berkaitan erat dengan bahasa *Natural Language Processing (NLP)* dan memerlukan langkah-langkah pra-pemrosesan agar bisa diklasifikasikan (Nurlayli and Nasichuddin 2019).

2.2 *Topik Modeling*

Topik modeling adalah teknik *dalam Natural Language Processing* yang fokus pada analisis algoritma teks (Bangkalang 2023). Metode ini menggunakan distribusi kata-kata untuk mengungkap pola tersembunyi dalam teks, seperti yang dilakukan oleh LSA, PLSA, dan terutama LDA yang sangat efektif dalam merangkum data teks besar berdasarkan topik kata-kata (Galuh Nurvinda K 2022).

2.3 *Data Colection*

Data Collection merupakan metode pengumpulan data dengan teknik *crawling*. *Crawling* adalah proses otomatis di mana program komputer, disebut sebagai *crawler* atau *spider*, mengumpulkan informasi dari berbagai situs web atau sumber data *online* secara sistematis (Galuh Nurvinda K 2022).

2.4 *Analisis Data*

Analisis data adalah metode untuk menggali sifat- sifat data dalam sistem informasi dengan tujuan memperdalam representasi, korelasi, dan signifikansinya (Paembonan and Abduh 2021).

2.5 *Preprocessing*

Preprocessing adalah langkah awal dalam mengklasifikasikan dokumen, mempersiapkan data agar terstruktur. Hasilnya berupa data numerik yang siap untuk diproses lebih lanjut (Patmawati and Yusuf 2021). Langkah awal dalam memproses dokumen adalah mengubahnya ke format yang memudahkan pencarian dokumen yang sesuai. Setiap tahapan dalam proses ini bertujuan membuat indeks dari dokumen-dokumen yang ada (Alfanzar and Rozas 2020).

1. *Text Cleaning*

Text Cleaning adalah langkah penting untuk membersihkan teks dari tanda baca, simbol, dan karakter non-huruf (Indrayuni 2019). Tahap *awal text cleaning* meliputi *casefolding*, penghapusan tanda baca, data duplikat, dan karakter tidak perlu, diikuti oleh cleaning lebih lanjut untuk menghilangkan atribut dan karakter tak diinginkan pada data.

2. *Stemming*

Stemming adalah upaya linguistik untuk mencari kata dasar tanpa memperhatikan artinya dengan menghilangkan awalan dan akhiran dari sebuah kata (Alfanzar and Rozas 2020). Pada langkah ini, kata-kata dipangkas atau dikembalikan ke bentuk dasarnya untuk keperluan sistem pencarian informasi, analisis teks, dan pemrosesan bahasa.

3. *Tokenisasi*

Tokenisasi adalah langkah penting dalam memproses teks di mana kata atau frasa dipisahkan menjadi unit-unit terpisah dalam sebuah *array* atau istilah (Alfanzar and Rozas 2020). Tahapan ini merubah kalimat-kalimat menjadi kata tunggal untuk digunakan dalam pemodelan topik.

4. *Stopword*

Stopword adalah langkah penting dalam *text preprocessing* untuk menghapus kata-kata umum yang sering muncul namun tidak banyak memberikan kontribusi pada pemahaman teks (Alfanzar and Rozas 2020).

Dalam pemodelan topik, langkah ini menghilangkan kata-kata yang tidak relevan atau tidak diinginkan terkait dengan kata kunci (Indrayuni 2019). Dalam membuat daftar kata yang akan digunakan dalam *stopword*, penulis melakukan dengan cara manual memilah kata yang tidak berkaitan dengan kata kunci. Daftar kata-kata yang dihasilkan diperoleh dari hasil *tokenisasi*.

2.6 *Text Processing*

Text preprocessing adalah langkah krusial untuk memperbaiki struktur data teks dengan cara melakukan *case folding, tokenizing, filtering*, dan agar formatnya lebih terstruktur (Tineges 2021).

1. Frekuensi Kata

Frekuensi kata merupakan jumlah kemunculan suatu kata dalam sebuah teks untuk menganalisis pola-pola, pemakaian, atau keterkaitan kata-kata dalam sebuah teks.

2. *WordCloud*

WordCloud adalah representasi visual kata-kata dalam teks di mana kata-kata yang sering muncul ditampilkan lebih besar, menciptakan gambaran secara grafis.

3. *PhraseCloud*

Pharse Cloud menunjukan keterkaitan setiap kata dengan kata yang lain. Setiap kata akan memiliki korelasi antara satu dengan yang lain.

4. Pasangan Kata

Dalam pemodelan, pasangan kata yang paling umum dapat mengungkapkan pola dan tren yang signifikan. Hhasilnya menggambarkan pemahaman yang mendalam tentang preferensi dan fokus dalam suatu model.

5. TF-IDF

TF-IDF adalah metode penilaian kata dalam sebuah dokumen dengan mempertimbangkan *frekuensi* kata tersebut dalam dokumen itu dan sebandingkan dengan frekuensi kata tersebut dalam seluruh koleksi dokumen (Deolika, Kusrini, and Luthfi 2019). Dalam TF-IDF, langkah awalnya adalah menghitung nilai TF untuk setiap kata dengan bobot awal 1. Sementara itu, nilai IDF dihitung dengan rumus tersendiri.

$$TF - IDF(word) = log\frac{td}{df} \quad (1)$$

TF-IDF mengukur pentingnya kata dalam koleksi dokumen dengan mempertimbangkan seberapa sering kata itu muncul dalam satu dokumen (TF) dan seberapa umumnya kata itu di seluruh dokumen (IDF).

6. *Latent Dirichlet Allocation (LDA)*

Latent Dirichlet Allocation (LDA) adalah teknik yang digunakan untuk mengidentifikasi topik-topik yang ada dalam suatu dataset serta untuk melakukan pemodelan terhadap topik-topik

tersebut (Patmawati and Yusuf 2021). Metode ini berupaya untuk mengelompokkan kata-kata yang sering muncul bersama dalam teks yang sama ke dalam topik yang sama (Santoso et al. 2022).

7. *NRE (Named Entity Recognition)*

NRE (Named Entity Recognition) teknik untuk mengidentifikasi entitas penting dari teks. *NRE Entity Frequensi* menggunakan teknik ini untuk menemukan entitas yang sering muncul dalam dokumen, membantu memahami topik utama dokumen tersebut.

2.7 *Klastering*

Klastering adalah metode pengelompokan data berdasarkan kesamaan karakteristik antara objek merupakan suatu proses di mana untuk mencari data dengan karakteristik yang serupa dan memberikan label yang tepat merupakan salah satu tantangan utama dalam aplikasi analisis data (Galuh Nurvinda K 2022). *Silhouette scores,* sering disimbolkan sebagai *Sscores*, digunakan untuk menghitung rata-rata perbedaan antara titik-titik dalam satu kluster dan antara kluster yang berbeda (Azizah, Widiharih, and Hakim 2022).

Metode Elbow adalah teknik untuk menentukan jumlah kluster optimal dengan mencari titik di mana penambahan kluster tidak signifikan lagi, terlihat seperti "siku" pada grafik (Ekasetya and Jananto 2020).

Adapun algoritma klastering yang digunakan yaitu:

1. K-Means

K-Means merupakan algoritma dalam *machine learning* yang digunakan untuk melakukan pengelompokan data ke dalam cluster berdasarkan pola atau kesamaan tertentu di antara data tersebut (Bryan Orleans 2022). Algoritma ini mengelompokkan data berdasarkan kesamaannya tanpa label. Tiap klaster memiliki *centroid* untuk mengurangi variasi data (Galuh Nurvinda K 2022). Data serupa dikelompokkan, yang berbeda akan dipisahkan. Kriteria kesamaannya bisa berupa jarak, pola, atau kepadatan (Paembonan and Abduh 2021).

2. K-Medoids

K-medoids adalah teknik klastering yang menggunakan titik data aktual sebagai pusat klaster, memilih medoid terdekat, dan memiliki kelebihan tahan terhadap pencilan, berguna untuk analisis pasar, bioinformatika, dan pemrosesan gambar meskipun membutuhkan komputasi-intensif pada data besar.

2.8 Evaluasi Hasil

Melakukan pemrosesan untuk mendapatkan hasil dari berbagai metode yang dilakukan dengan cara evaluasi hasil dengan metode sebagai berikut:

1. *SSE (Sum of Squared Errors)*

Selain itu, *SSE (Sum of Squared Errors)* digunakan untuk menilai sejauh mana model sesuai dengan data, menghitung kesalahan kuadrat antara nilai prediksi dan nilai sebenarnya. Tujuan utamanya yaitu meminimalkan nilai SSE untuk menjelaskan variasi data secara efektif dan membandingkan kinerja model, dengan rumus 3.

$$SSE = \sum_{i-1}^{n}(y_1 - y)2 \quad (3)$$

Di mana yiadalah nilai aktual dari titik data ke-i, dan $\hat{y}i$ adalah nilai yang diprediksi oleh model untuk titik data ke-i.

2. *Calinski Harabasz Score*

Calinski harabasz score merupakan metode untuk mengukur jumlah kuadrat antar klaster dan di dalam klaster, sementara jumlah klaser (K) ditentukan dengan menggunakan nilai maksimum dari skor *calinski-harabasz CHK*. Berikut rumus dari *calinski harabasz*:

$$CH_K = \frac{SSB}{SSW} \times \frac{N-K}{K-1} \quad (2)$$

Di mana K adalah jumlah klaster, SSB adalah total jumlah kuadrat jarak antar klaster, dan SSW adalah total jumlah kuadrat jarak di dalam klaster (Azizah, Widiharih, and Hakim 2022).

3. Hasil dan Pembahasan

Penelitian ini menyusun kategori-kategori topik yang muncul pada website kemenkes tahun 2021-2023 dengan jumlah data sebanyak 1498 baris data dengan 1 kolom. Data diambil dengan metode crawling menggunakan ekstensi Instan Data Scraper dari peramban *(Google Chrome)*. Berikut merupakan tautan terkait berita kesehatan kemenkes Indonesia yang dipakai: https://sehatnegeriku.kemkes.go.id/topik/rilis-media/.

3.1 Analisis Data

Pada penelitian ini dilakukan analisis data dengan berbagai metode untuk mencapai hasil yang diinginkan, dimulai dengan pemeriksaan data, pengolahan data, penghapusan atribut yang tidak penting dan data duplikat, serta membersihkan teks sebagai bagian dari tahapan *preprocessing*.

ISSN: 2541-1004
e-ISSN: 2622-4615
https://doi.org/10.32493/informatika.v9i2.38971

```
                                                   Text
0     Kemenkes Raih 3 Penghargaan dalam Ajang TOP DI...
1     Butuh Kualifikasi Dokter Tinggi, Menkes Minta ...
2     RSUP dr. Ben Mboi Diresmikan Presiden, Warga N...
3     Pabrik Fraksionasi Plasma Pertama di Indonesia...
4     Menkes Kukuhkan Tenaga Cadangan Kesehatan Tipe...
...                                                 ...
1493           Kenali Gejala Stroke dengan Metode FAST
1494  Hasil Penyelidikan Kepolisian Tidak Ditemukan ...
1495  Pemerintah Manfaatkan Momentum Penurunan Kasus...
1496                      Indonesia Waspadai Varian Mu
1497  Pusat Kesehatan Haji Kemenkes Gelar Sosialisas...

[1498 rows x 1 columns]
Jumlah Data: 1498
Jumlah Baris: 1498, Jumlah Kolom: 1
Jumlah Missing Value: 0
Jumlah Data Duplikat: 151
Persentase Data yang Akan Dihapus: 10.08%

Jumlah Total Data Sebelum Pembersihan: 1498
```

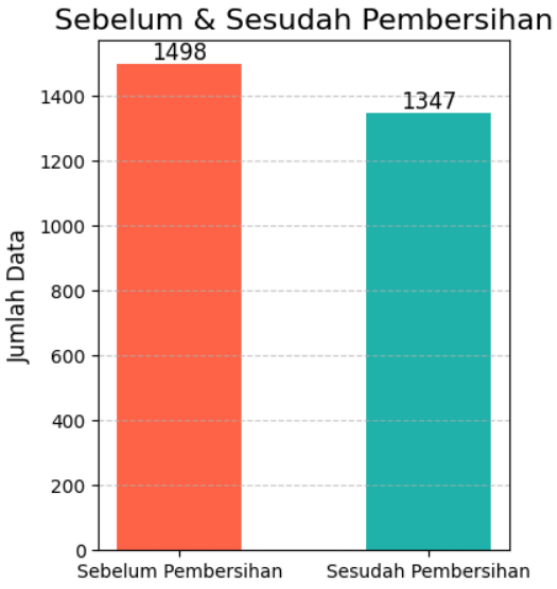
Gambar 2. Dataset Berita Kemenkes

Pada dataset ini memiliki jumlah data sebanyak 1498 data baris, dan 1 data kolom dengan label text. *Missing value* tidak ada, dengan jumlah data duplikat 151 data. Persentase data yang akan dihapus sebesar 10.08%.

3.2 *Preprocessing*

Pada tahapan ini dilakukan berbagai macam cara preprocessing data sebelum melanjutkan ke tapahpan *text processing*. Adapun tahapan-tahapan itu antara lain:

1. Text Cleaning



Gambar 3. Grafik Data Sebelum dan sesudah Pembersihan

Berikut di atas merupakan perbandingan jumlah kata sebelum dan sesudah dilakukan cleaning data. Pada proses ini menghasilkan data dengan jumlah data sebanyak 1347 data yang sebelumnya sebanyak 1498 data.

2. *Stemming*

```
                                                Text
0  kemenkes raih harga dalam ajang top digital aw...
1  butuh kualifikasi dokter tinggi menkes minta r...
2  rsup dr ben mboi resmi presiden warga ntt tida...
3  pabrik fraksionasi plasma pertama di indonesia...
4  menkes kukuh tenaga cadang sehat tipe target d...
```

Gambar 4. Hasil Stamming

Pada tahapan ini dilakukan pembersihan pada kata yang memiliki imbuhan kata menjadi kata dasar dan menyeragamkan huruf menjadi huruf kecil. Seperti kata penghargaan menjadi harga, kukuhkan menjadi kukuh dan lain sebagainya.

3. *Tokenisasi*

```
                                                Text
0  [kemenkes, raih, harga, dalam, ajang, top, dig...
1  [butuh, kualifikasi, dokter, tinggi, menkes, m...
2  [rsup, dr, ben, mboi, resmi, presiden, warga, ...
3  [pabrik, fraksionasi, plasma, pertama, di, ind...
4  [menkes, kukuh, tenaga, cadang, sehat, tipe, t...
```

Gambar 5. Hasil Tokenisasi

Pada tahapan ini melakukan proses merubah kalimat menjadi ke dalam bentuk kata dengan memisahkan kata dengan tanda koma pada setiap kata seperti kemenkes, raih, harga, dalam, ajang, top, dan lain sebagainya.

4. *Stopword*

```
                                Text
0                          [kemenkes]
1  [dokter, dr, prioritas, dokter]
2      [dr, presiden, warga, obat]
3              [plasma, indonesia]
4                     [sehat, who]
```

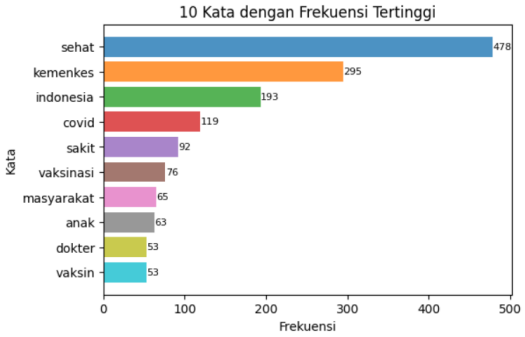
Gambar 6. Hasil Stopword

Tahapan ini menghapus kata-kata yang tidak berkaitan dengan kata kunci "Kesehatan-Kemenkes" akan dilakukan penghapusan, sehingga menghasilkan kata yang berkaitan dengan topik judul pembahasan. Daftar kata yang digunakan untuk *stopword* dibuat secara manual dengan memilah kata yang berkaitan dengan kata kunci.

ISSN: 2541-1004
e-ISSN: 2622-4615
https://doi.org/10.32493/informatika.v9i2.38971

3.3 *Text Processing*

1. Frekuensi Kata



Gambar 7. Frekuensi Kata

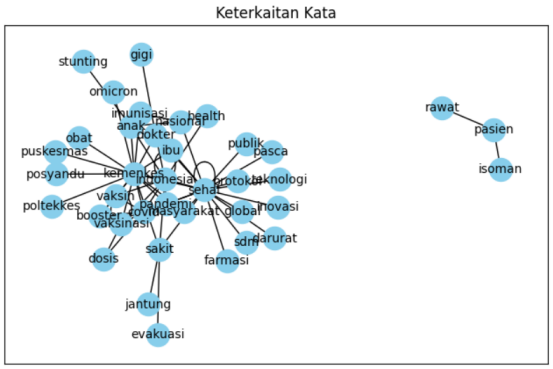
Pada frekuensi kata dengan jumlah frekuensi tertinggi dari 10teratas yaitu sehat, kemenkes, Indonesia, covid, sakit, vaksiansi, masyarakat, anak, dokter dan vaksin.

2. *WordCloud*

Gambar 8. Wordcloud

Pada *wordcloud* kata yang sering dibahas pada berita kemenkes yaiti sehat, kemenkes, dan Indonesia.

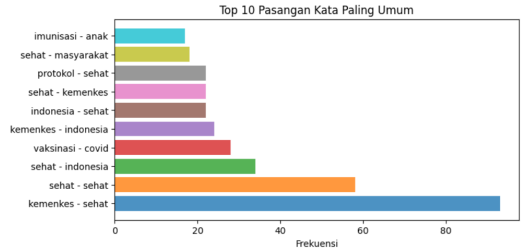
3. *Pharse Cloud*



Gambar 9. Pharse Cloud

Pada *pharse cloud* membahas keterkaitan kata satu dengan yang lainya. Seperti pasien bekaitan dengan rawat dan isoman.

4. Top 10 Pasangan Kata Paling Umum



Gambar 10. Pasangan Kata

Pasangan kata yang sering muncul yaitu imunisasi- anak, sehat-masyarakat, protocol-sehat, sehat- kemenkes, Indonesia-sehat, vaksinasi-covid, sehat- indonesia, sehat-sehat, dan kemenkes-sehat.

5. *TF-IDF Feature*

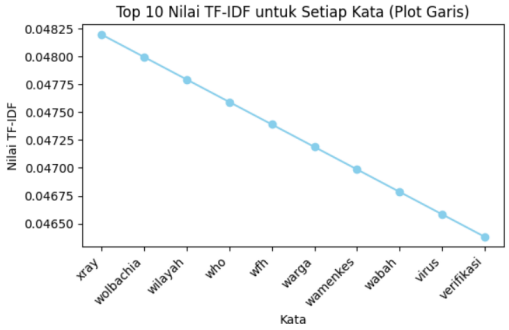
TF-IDF Features:

	No	Kata	Frekuensi	TF	IDF	TF-IDF
0	1	adaptasi	2.0	0.000070	5.783316	0.000403
1	2	ahli	3.0	0.000105	5.783316	0.000605
2	3	aid	4.0	0.000139	5.783316	0.000807
3	4	alkes	5.0	0.000174	5.783316	0.001008
4	5	anak	6.0	0.000209	5.783316	0.001210
..	...	...	...	...	...	...
233	234	wfh	235.0	0.008194	5.783316	0.047389
234	235	who	236.0	0.008229	5.783316	0.047591
235	236	wilayah	237.0	0.008264	5.783316	0.047793
236	237	wolbachia	238.0	0.008299	5.783316	0.047994
237	238	xray	239.0	0.008334	5.783316	0.048196

[238 rows x 6 columns]

Gambar 11. TF-IDF

Pada tf-idf kata dengan nilai frkuensi tertinggi yaitu pada kata xray dengan nilai frekuensi 239.0 sedangkan nilai *frekuensi* terendah pada kata adaptasi dengan nilai 2.0.



Gambar 12. Grafik TF-IDF

Grafik diatas menunjukan penurunan yang begitu signifikan pada 10 kata tertinggipada *frekuensi tf-idf* dengan rentan nilai dari 0.04825 – 0.04650.

6. LDA Model

```
LatentDirichletAllocation
LatentDirichletAllocation(n_components=5, random_state=1)
```

Gambar 13. Woercloud Topik Model 1

ISSN: 2541-1004
e-ISSN: 2622-4615
https://doi.org/10.32493/informatika.v9i2.38971

Pada *LDA model* ini menampilkan sebanyak 5 topik yang sering dibahas pada laman website kemenkes dalam kurun waktu dari tahun 2021-2023 yaitu:

Gambar 14. Woercloud Topik Model 1

Topik 1 membahas berkaitan dengan Kesehatan masyarakat Indonesia terkait terhadap kanker, tbc, dan covid.

Gambar 15. Woercloud Topik Model 2

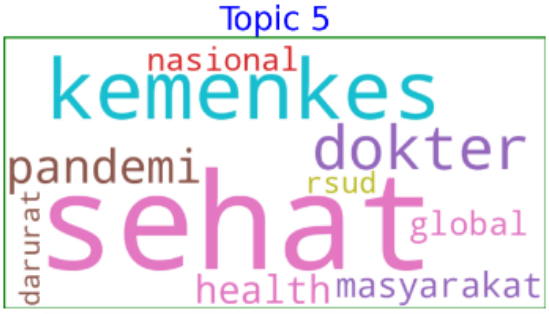
Topik 2 membahas berkaitan dengan vaksinasi covid dan imunasisi dengan kemungkinan komlikasi terhadap ginjal dan jantung.

Gambar 16. Woercloud Topik Model 3

Pada topik 3 masih membahas berkaitan dengan vaksinasi covid hanya saja pada topik ini membahas berkaitan inofasi yang dilakukan wamenkes terhadap pelayan vaksinasi dan penggunaan aplikasi pedulilindungi.

Gambar 17. Woercloud Topik Model 4

Pada topik 4 membahas berkaitan dengan pelayanan posyandu yang dinaungi puskesmas dan kemenkes terhadap pelayanan imunisasi, stunting, gizi, dan penangan omicron.



Gambar 18. Woercloud Topik Model 5

Pada topik 5 membahas berkaitan dengan Kesehatan global setelah adanya pandemic covid, kemenkes melakukan himbauan agar rumah sakit dan para dokter agar untuk melakukan pemantauan terhadap masyarakat agar tercapainya masyarakat yang sehat.

7. NRE Entity Frequency

```
Tabel NRE Entity Frequency:
                                  Entity  Frequency
0                        (indonesia, GPE)         10
1          (covid deteksi tbc, PERSON)         10
2             (jantung imunisasi, ORG)         10
3                        (vaksin, GPE)         10
4  (pedulilindungi inovasi teknologi, PERSON)         10

Topik Model Banyak Membahas Tentang: ('indonesia', 'GPE')
```

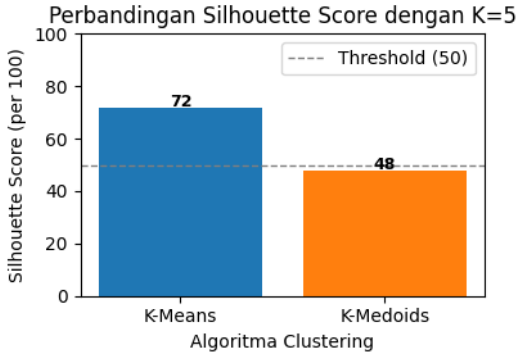
Gambar 19. Entity Frekuensi

Topik model banyak membahas, salah satunya tentang: (‘Indonesia’,‘GPE’) (*Geo Political Entity*) merujuk kepada entitas *geografis* atau politik, seperti negara, kota, atau wilayah administratif.

ISSN: 2541-1004
e-ISSN: 2622-4615
https://doi.org/10.32493/informatika.v9i2.38971

3.4 Klastering

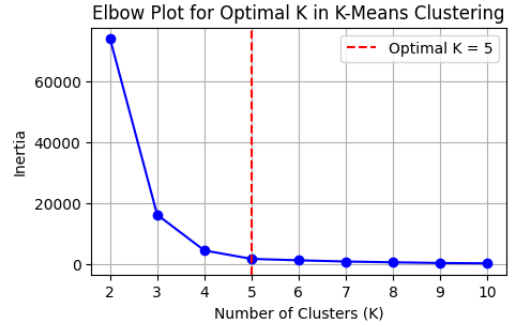
1. Silhoutte Score



Gambar 20. Perbandingan Algoritma

Perbandingan ini menggunakan *silhouette score* menggunakan k=5 dengan *threshold* 50%. Menghasilkan akurasi K-Means sebesar 72% dan k-medoids sebesar 48%.

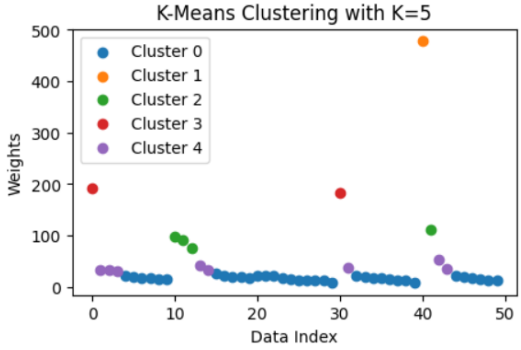
2. Elbow Method



Gambar 20. Elbow Method

Dalam menetukan jumlah klaster yang optimal dalam k-mean menggunaan grafik *elbow method* menghasilkan penurunan nilai yang signifikan pada k=5, oleh karena itu pada data ini cocok menggunakan k=5 untuk menetukan jumlah klaster pada *k-mean clustering* dengan nilai inertia 1676.

3. K-Means Clustering



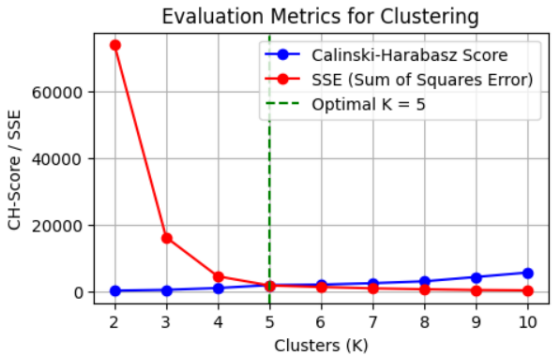
Gambar 21. Klustering K-Means

Hasil klustering dengan menggunakan algoritma K-Means pada data hasil topik modeling menghasilkan sebaran data sebanyak 5 klaster yaitu:

Gambar 22. Hasil Klustering

Klaster 0 berkaitan dengan health, darurat, tbc, pasien, teknologi, deteksi, anak, wamenkes, jantung, ginjal, protocol, covid, kanker, dunia, masyarakat. omicron, vaksinasi, dan imunisasi. Klaster 1 berkaitan dengan kata sehat. Klaster 2 berkaitan dengan sakit, kemenkes, covid, dan vaksinasi. Klaster 3 berkaitan dengan kata Indonesia, dan kemenkes. Klaster 4 berkaitan dengan dokter, anak, obat, masyarakat, pasien, stunting, pandemic, dan boster.

3.5 Evaluasi Calinski Harabasz Score & SSE (Sum of Squared Errors)



Gambar 23. CH Score dan SSE

Hasil evaluasi *Sum of Squares Error (SSE)* dengan k=5 menunjukkan penurunan yang signifikan dengan nilai metrik SSE berkisar antara (75.000-4.500). Dengan penurunan nilai yang signifikan ini menunjukan data cluster semakin seragam pesebaran datanya dan semakin kecil nilai *error* pada data cluster.

Sementara itu, *Calinski Harabasz Score* dengan k=5 menunjukkan stabilitas nilai yang tinggi, dengan metrik *CH Score* menghasilkan nilai antara (0-7.000). Dengan meningkatnya nilai pada data makan semakin baik data pada cluster.

4. Kesimpulan

Dari hasil penelitian ini, dapat disimpulkan bahwa tingkat akurasi pada pemodelan dan klasterisasi sangat memuaskan, dengan algoritma K-Means mencapai akurasi sebesar 72%, melebihi ambang batas yang telah ditetapkan sebelumnya sebesar 50%. Hasil evaluasi *Sum of Squares Error* (SSE) dengan k=5 menunjukkan penurunan yang signifikan, dengan nilai metrik SSE berkisar antara 75.000 hingga 4.500. Penurunan nilai yang signifikan ini mengindikasikan bahwa data cluster semakin seragam dalam pesebarannya, dan semakin kecil nilai error pada data cluster. Sementara itu, *Calinski Harabasz Score* dengan k=5 menunjukkan stabilitas nilai yang tinggi, dengan metrik CH *Score* menghasilkan nilai antara 0 hingga 7.000. Meningkatnya nilai pada metrik ini menunjukkan bahwa data dalam cluster semakin baik.

Sebelum penelitian ini dilakukan, konten pada situs web Kementerian Kesehatan (Kemenkes) tersebar tanpa struktur yang jelas, membuat pengguna kesulitan menemukan informasi kesehatan yang mereka butuhkan dengan cepat dan efisien. Hal ini mengakibatkan penurunan kualitas pengalaman pengguna dan potensi penurunan kepercayaan terhadap sumber informasi kesehatan resmi. Setelah penelitian ini diselesaikan, dampak utama yang dihasilkan adalah peningkatan signifikan dalam pengelompokan dan aksesibilitas informasi di situs web tersebut. Dengan penerapan algoritma K-Means, konten kini dikelompokkan secara tematis ke dalam lima klaster utama, memungkinkan pengguna untuk *menavigasi* dan menemukan informasi yang relevan dengan lebih mudah. Hasilnya, kualitas penyebaran informasi kesehatan meningkat, mendukung pengambilan keputusan berbasis data yang lebih tepat, serta memperbaiki pengalaman pengguna secara keseluruhan. Dampak ini tidak hanya mempermudah akses informasi, tetapi juga memperkuat peran situs web Kemenkes sebagai sumber informasi kesehatan yang handal dan efisien.

Referensi

Alfanzar, Alif Iffan, and Indri Sudanawati Rozas. 2020. "Topic Modelling Skripsi Menggunakan Metode Latent." *JSiI (Jurnal Sistem Informasi)* 7(1): 7–13.

Anjar, Agus, Muhammad Khairul Ritonga, and Toni Toni. 2021. "DAMPAK POSITIF DAN NEGATIF PERKEMBANGAN TEKNOLOGI KOMUNIKASI TERHADAP MAHASISWA PPKn FKIP LABUHANBATU." *Civitas (Jurnal Pembelajaran Dan Ilmu Civic)* 7(2): 41–44.

Azizah, Anestasya Nur, Tatik Widiharih, and Arief Rachman Hakim. 2022. "Kernel K-Means Clustering Untuk Pengelompokan Sungai Di Kota Semarang Berdasarkan Faktor Pencemaran Air." *Jurnal Gaussian* 11(2): 228–36.

Bangkalang, Dwi Hosanna. 2023. "Analisis Dan Penerapan Topic Modeling Pada Judul Tugas Akhir Mahasiswa Menggunakan Metode Latent Dirichlet Allocation (Lda)." 8(4): 1275–87.

Bryan Orleans, Edi Purnomo Putra. 2022. "Clustering Algoritma (K-Means)." *Binus*. https://sis.binus.ac.id/2022/01/31/clustering-algoritma-k-means/.

Cahyono, Nuri, and Angga Reni Dwi Astuti. 2023. "Analisis Topic Modelling Persepsi Pengguna Internet Menggunakan Metode Latent Dirichlet Allocation." *Indonesian Journal of Computer Science* 12(1): 326–34.

Deolika, Agatha, Kusrini Kusrini, and Emha Taufiq Luthfi. 2019. "Analisis Pembobotan Kata Pada Klasifikasi Text Mining." *Jurnal Teknologi Informasi* 3(2): 179.

Dinda Adimanggala, Fitra Abdurrachman Bachtiar, and Eko Setiawan. 2021. "Evaluasi Topik Tersembunyi Berdasarkan Aspect Extraction Menggunakan Pengembangan Latent Dirichlet Allocation." *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* 5(3): 511–19.

Doni. 2022. "Literasi Digital Masyarakat Indonesia Membaik." *kominfo*. https://www.kominfo.go.id/content/detail/39858/literasi-digital-masyarakat-indonesia-membaik/0/artikel.

Ekasetya, Vada Annisa, and Arief Jananto. 2020. "Klusterisasi Optimal Dengan Elbow Method Untuk Pengelompokan Data Kecelakaan Lalu Lintas Di Kota Semarang." *Jurnal Dinamika Informatika* 12(1): 20–28.

Galuh Nurvinda K. 2022. "Algoritma Clustering Data Science Terupdate 2022." *DQLab*. https://dqlab.id/algoritma-clustering-data-science-terupdate-2022.

Indrayuni, Elly. 2019. "Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes." *Jurnal Khatulistiwa Informatika*

7(1): 29–36.

Nurlayli, Akhsin, and Moch. Ari Nasichuddin. 2019. "Topik Modeling Penelitian Dosen Jptei Uny Pada Google Scholar Menggunakan Latent Dirichlet Allocation." *Elinvo (Electronics, Informatics, and Vocational Education)* 4(2): 154–61.

Paembonan, Solmin, and Hisma Abduh. 2021. "Penerapan Metode Silhouette Coefficient Untuk Evaluasi Clustering Obat." *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik* 6(2): 48.

Patmawati, Patmawati, and Muhammad Yusuf. 2021. "Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter Oleh Pejabat Negara." *Building of Informatics, Technology and Science (BITS)* 3(3): 122–29.

Rusdhi, Vira Faradhiba, and Ilmiyati Sari. 2022. "Identifikasi Topik Artikel Berita Menggunakan Topic Modelling Dengan Latent Dirichlet Allocation." *Jurnal Ilmiah Informatika Komputer* 27(2): 169–76.

Santoso, Kevin Rafi Adjie Putra, Asmaul Husna, Nadia Widyawati Putri, and Nur Aini Rakhmawati. 2022. "Analisis Topik Tagar Covidindonesia Pada Instagram Menggunakan Latent Dirichlet Allocation." *JISKA (Jurnal Informatika Sunan Kalijaga)* 7(1): 1–9.

Tineges, Rian. 2021. "Tahapan Text Preprocessing Dalam Teknik Pengolahan Data." *DQLab*. https://dqlab.id/tahapan-text-preprocessing-dalam-teknik-pengolahan-data.

Wahyudin. 2020. "Aplikasi Topic Modeling Pada Pemberitaan." *Seminar Nasional Official Statistic: Pengembangan Official Statistics dalam mendukung Implementasi SDG's*: 309–18.

Woro, Verianty Anjar. 2022. "Yang Dimaksud Dengan Berita Adalah Informasi Peristiwa, Kenali Unsur Dan Faktornya." *Lioutan6.com*. https://www.liputan6.com/hot/read/5142500/yang-dimaksud-dengan-berita-adalah-informasi-peristiwa-kenali-unsur-dan-faktornya?page=4 (December 21, 2023).